Graphonomy: Universal Human Parsing via Graph Transfer Learning

Ke Gong[1,2†], Yiming Gao[1†], Xiaodan Liang[1*],
Xiaohui Shen[3], Meng Wang[4], Liang Lin[1,2]

[1]Sun Yat-sen University [2]DarkMatter AI Research [3]ByteDance AI Lab [4]Hefei University of Technology

kegong936@gmail.com, gaoym9@mail2.sysu.edu.cn, xdliang328@gmail.com,
shenxiaohui@gmail.com, wangmeng@hfut.edu.cn, linliang@ieee.org

Abstract

Prior highly-tuned human parsing models tend to fit towards each dataset in a specific domain or with discrepant label granularity, and can hardly be adapted to other human parsing tasks without extensive re-training. In this paper, we aim to learn a ***single*** *universal human parsing model that can tackle all kinds of human parsing needs by unifying label annotations from different domains or at various levels of granularity. This poses many fundamental learning challenges,* e.g. *discovering underlying semantic structures among different label granularity, performing proper transfer learning across different image domains, and identifying and utilizing label redundancies across related tasks.*

*To address these challenges, we propose a new universal human parsing agent, named "Graphonomy", which incorporates hierarchical graph transfer learning upon the conventional parsing network to encode the underlying label semantic structures and propagate relevant semantic information. In particular, Graphonomy first learns and propagates compact high-level graph representation among the labels within one dataset via Intra-Graph Reasoning, and then transfers semantic information across multiple datasets via Inter-Graph Transfer. Various graph transfer dependencies (*e.g.*, similarity, linguistic knowledge) between different datasets are analyzed and encoded to enhance graph transfer capability. By distilling universal semantic graph representation to each specific task, Graphonomy is able to predict all levels of parsing labels in one system without piling up the complexity. Experimental results show Graphonomy effectively achieves the state-of-the-art results on three human parsing benchmarks as well as advantageous universal human parsing performance.*

1. Introduction

Human visual systems are capable of accomplishing holistic human understanding at a single glance on a person image, e.g., separating the person from the background, understanding the pose, and recognizing the clothes the person wears. Nevertheless, recent research efforts on human understanding have been devoted to developing numerous highly-specific and distinct models for each individual application, e.g. foreground human segmentation task [8, 15], coarse clothes segmentation task [25, 28] and fine-grained human part/clothes parsing task [14, 39]. Despite the common underlying human structure and shared intrinsic se-

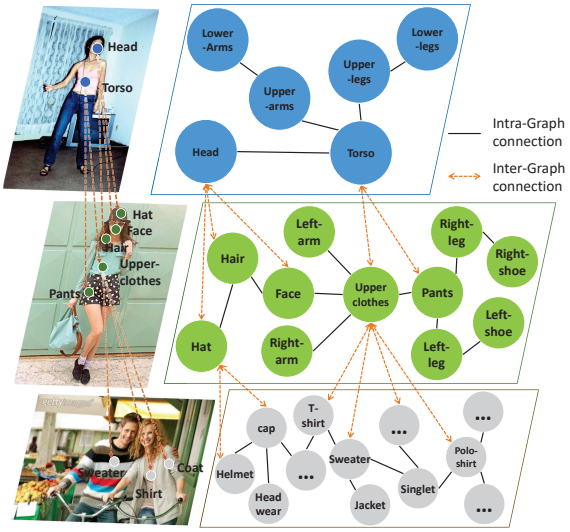

Figure 1. With huge different granularity and quantity of semantic labels, human parsing is isolated into multiple level tasks that hinder the model generation capability and data annotation utilization. For example, the *head* region on a dataset is further annotated into several fine-grained concepts on another dataset, such as *hat*, *hair* and *face*. However, different semantic parts still have some intrinsic and hierarchical relations (*e.g.*, *Head* includes the *face*. *Face* is next to *hair*), which can be encoding as intra-graph and inter-graph connections for better information propagation. To alleviate the label discrepancy issue and take advantage of their semantic correlations, we introduce a universal human parsing agent, named as "Graphonomy", which models the global semantic coherency in multiple domains via graph transfer learning to achieve multiple levels of human parsing tasks.

† Equal contribution. *Corresponding Author.

mantic information (e.g. upper-clothes can be interpreted as coat or shirt), these highly-tuned networks have sacrificed the generalization capability by only fitting towards each dataset domain and discrepant label granularity. It is difficult to directly adapt the model trained on one dataset to another related task, and thus requires redundant heavy data annotation and extensive computation to train each specific model. To address these realistic challenges and avoid training redundant models for correlated tasks, we make the first attempt to investigate a *single* universal human parsing agent that tackles human parsing tasks at different coarse to fine-grained levels, as illustrated in Fig. 1.

The most straightforward solution to universal human parsing would be posing it as a multi-task learning problem, and integrating multiple segmentation branches upon one shared backbone network [2, 14, 22, 25, 28]. This line of research only considers the brute-force feature-level information sharing while disregarding the underlying common semantic knowledge, such as label hierarchy, label visual similarity, and linguistic/context correlations. More recently, some techniques are explored to capture the human structure information by resorting to complex graphical models (*e.g.*, Conditional Random Fields (CRFs)) [2], self-supervised loss [14] or human pose priors [9, 12, 23]. However, they did not explicitly model the semantic correlations of different body parts and clothing accessories, and still show unsatisfactory results for rare fine-grained labels.

One key factor of designing a universal human parsing agent is to have proper transfer learning and knowledge integration among different human parsing tasks, as the label discrepancy across different datasets [6, 13, 14, 39] largely hinders direct data and model unification. In this paper, we achieve this goal by explicitly incorporating human knowledge and label taxonomy into intermediate graph representation learning beyond local convolutions, called "Graphonomy" (graph taxonomy). Our Graphonomy learns the global and common semantic coherency in multiple domains via graph transfer learning to solve multiple levels of human parsing tasks and enforce their mutual benefits upon each other.

Taking advantage of geometric deep learning [19, 20], our Graphonomy simply integrates two cooperative modules for graph transfer learning. First, we introduce Intra-Graph Reasoning to progressively refine graph representations within the same graph structure, in which each graph node is responsible for segmenting out regions of one semantic part in a dataset. Specifically, we first project the extracted image features into a graph, where pixels with similar features are assigned to the same semantic vertex. We elaborately design the adjacency matrix to encode the semantic relations, constrained by the connection of human body structure, as shown in Fig. 3. After the message propagation via graph convolutions, the updated vertexes are re-projected to make the visual feature maps more discriminative for pixel-level classification.

Additionally, we build an Inter-Graph Transfer module to attentively distill related semantics from the graph in one domain/task to the one in another domain, which bridges the semantic labels from different datasets, and effectively utilize the annotations at multiple levels. To enhance graph transfer capability, we make the first effort to exploit various graph transfer dependencies among different datasets. We encode the relationships between two semantic vertexes from different graphs by computing their feature similarity as well as the semantic similarity encapsulated with linguistic knowledge.

We conduct experiments on three human parsing benchmarks that contain diverse semantic body parts and clothes. The experimental results show that by seamlessly propagating information via Intra-Graph Reasoning and Inter-Graph Transfer, our Graphonomy is able to associate and distill high-level semantic graph representation constructed from different datasets, which effectively improves multiple levels of human parsing tasks.

Our contributions are summarized in the following aspects. 1) We make the first attempts to tackle all levels of human parsing tasks using a single universal model. In particular, we introduce Graphonomy, a new Universal Human Parsing agent that incorporates hierarchical graph transfer learning upon the conventional parsing network to predict all labels in one system without piling up the complexity. 2) We explore various graph transfer dependencies to enrich graph transfer capability, which enables our Graphonomy to distill universal semantic graph representation and enhance individualized representation for each label graph. 3) We demonstrate the effectiveness of Graphonomy on universal human parsing, showing that it achieves the state-of-the-art results on three human parsing datasets.

2. Related Work

Human Parsing. Human parsing has recently attracted a huge amount of interests and achieved great progress with the advance of deep convolutional neural networks and large-scale datasets. Most of the prior works focus on developing new structures and auxiliary information guidance to improve general feature representation, such as dilated convolution [2, 38], LSTM structure [24, 26, 27], encoder-decoder architecture [3], and human pose constraints [12, 23, 36]. Although these methods show promising results on each human parsing dataset, they directly use one flat prediction layer to classify all labels, which disregards the intrinsic semantic correlations across concepts and utilize the annotations in an inefficient way. Moreover, the trained model cannot be directly applied to another related task without heavy fine-tuning. In this paper, we investigate universal human parsing via graph transfer learning,

where each graph encodes a set of concepts in the taxonomy, and all graphs constructed from different datasets are connected following the transfer dependencies to enforce semantic feature propagation.

Multi-task Learning. Aiming at developing systems that can provide multiple outputs simultaneously for an input, multi-task learning has experienced great progress [8, 10, 13, 23, 36, 37]. For example, Gong *et al*. [13] jointly optimized semantic part segmentation and instance-aware edge detection in an end-to-end way and makes these two correlated tasks mutually beneficial. Xiao *et al*. [37] introduced a multi-task network and training strategy to handle heterogeneous annotations for unified perceptual scene parsing. However, these approaches simply create several branches for different tasks respectively, without exploring explicit relationships among the correlated tasks. In contrast to the existing multi-task learning pipelines, we explicitly model the relations among different label sets and extract a unified structure for universal human parsing via graph transfer learning.

Knowledge-guided Graph Reasoning. Many research efforts recently model domain knowledge as a graph for mining correlations among labels or objects in images, which has been proved effective in many tasks [5, 19, 20, 29, 35]. For example, Chen *et al*. [5] leveraged local region-based reasoning and global reasoning to facilitate object detection. Liang *et al*. [29] explicitly constructed a semantic neuron graph network by incorporating the semantic concept hierarchy. On the other hand, there are some sequential reasoning models for relationships [4, 21]. In these works, a fixed graph is usually considered, while our Graphonomy makes further efforts from external knowledge embedding to graph representation transfer.

Transfer Learning. Our approach is also related to transfer learning [32], which bridges different domains or tasks to mitigate the burden of manual labeling. LSDA [17] transformed whole-image classification parameters into object detection parameters through a domain adaptation procedure. Hu *et al*. [18] considered transferring knowledge learned from bounding box detection to instance segmentation. Our method transfers high-level graph representations in order to reduce the label discrepancy across different datasets.

3. Graphonomy

In order to unify all kinds of label annotations from different resources and tackle different levels of human parsing needs in one system, we aim at explicitly incorporating hierarchical graph transfer learning upon the conventional parsing network to compose a universal human parsing model, named as Graphonomy. Fig. 2 gives an overview of our proposed framework. Our approach can be embedded in any modern human parsing system by enhancing its original image features via graph transfer learning. We first learn and propagate compact high-level semantic graph representation within one dataset via Intra-Graph Reasoning, and then transfer and fuse the semantic information across multiple datasets via Inter-Graph Transfer driven by explicit hierarchical semantic label structures.

3.1. Intra-Graph Reasoning

Given local feature tensors from convolution layers, we introduce Intra-Graph Reasoning to enhance local features, by leveraging global graph reasoning with external structured knowledge. To construct the graph, we first summarize the extracted image features into high-level representations of graph nodes. The visual features that are correlated to a specific semantic part (*e.g*., *face*) are aggregated to depict the characteristic of its corresponding graph node.

Firstly, We define an undirected graph as $G = (V, E)$ where V denotes the vertices, E denotes the edges, and $N = |V|$. Formally, we use the feature maps $X \in \mathbb{R}^{H\times W\times C}$ as the module inputs, where H, W and C are height, width and channel number of the feature maps. We first produce high-level graph representation $Z \in \mathbb{R}^{N\times D}$ of all N vertices, where D is the desired feature dimension for each $v \in V$, and the number of nodes N typically corresponds to the number of target part labels of a dataset. Thus, the projection can be formulated as the function ϕ:

$$Z = \phi(X, W), \tag{1}$$

where W is the trainable transformation matrix for converting each image feature $x_i \in X$ into the dimension D.

Based on the high-level graph feature Z, we leverage semantic constraints from the human body structured knowledge to evolve global representations by graph reasoning. We introduce the connections between the human body parts to encode the relationship between two nodes, as shown in Fig 3. For example, *hair* usually appears with the *face* so these two nodes are linked. While the *hat* node and the *leg* node are disconnected because they have nothing related.

Following Graph Convolution [19], we perform graph propagation over representations Z of all part nodes with matrix multiplication, resulting in the evolved features Z^e:

$$Z^e = \sigma(A^e Z W^e), \tag{2}$$

where $W^e \in \mathbb{R}^{D\times D}$ is a trainable weight matrix and σ is a nonlinear function. The node adjacency weight $a_{v\to v\prime} \in A^e$ is defined according the edge connections in $(v, v\prime) \in E$, which is a normalized symmetric adjacency matrix. To sufficiently propagate the global information, we employ such graph convolution multiple times (3 times in practice).

Finally, the evolved global context can be used to further boost the capability of image representation. Similar

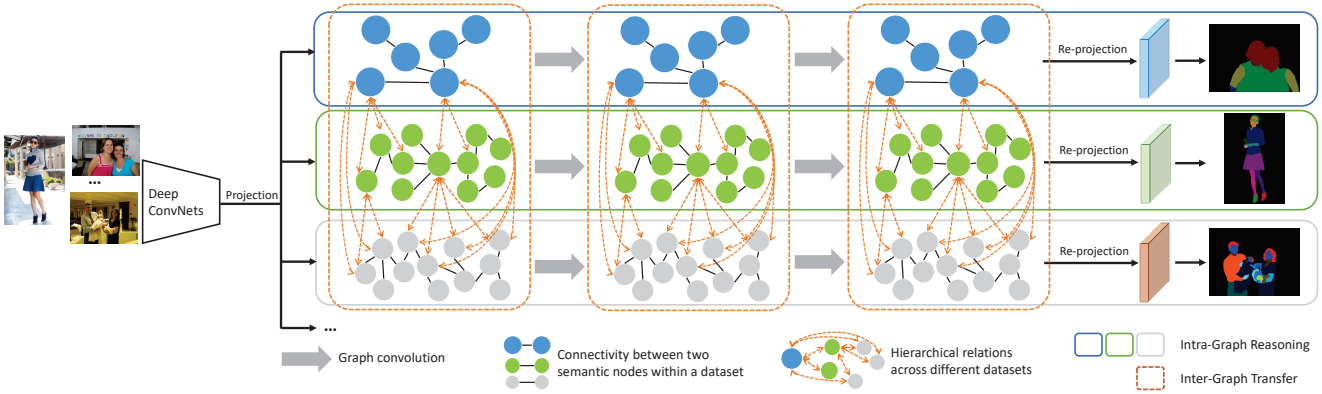

Figure 2. Illustration of our Graphonomy that tackles universal human parsing via graph transfer learning to achieve multiple levels of human parsing tasks and better annotation utilization. The image features extracted by deep convolutional networks are projected into a high-level graph representation with semantic nodes and edges defined according to the body structure. The global information is propagated via Intra-Graph Reasoning and re-projected to enhance the discriminability of visual features. Further, we transfer and fuse the semantic graph representations via Inter-Graph Transfer driven by hierarchical label correlation to alleviate the label discrepancy across different datasets. During training, our Graphonomy takes advantage of annotated data with different granularity. For inference, our universal human parsing agent generates different levels of human parsing results taking an arbitrary image as input.

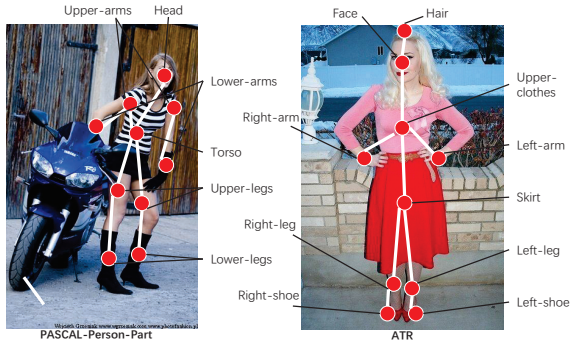

Figure 3. Examples of the definite connections between each two human body parts, which is the foundation to encode the relations between two semantic nodes in the graph for reasoning. Two nodes are defined related if they are connected by a white line.

to the projection operation (Eq. 1), we again use another transformation matrix to re-project the graph nodes to images features. We apply residual connection [16] to further enhance visual representation with the original feature maps X. As a result, The image features are updated by the weighted mappings from each graph node that represents different characteristics of semantic parts.

3.2. Inter-Graph Transfer

To attentively distill relevant semantics from one source graph to another target graph, we introduce Inter-Graph Transfer to bridge all semantic labels from different datasets. Although different levels of human parsing tasks have diverse distinct part labels, there are explicit hierarchical correlations among them to be exploited. For example, *torso* label in a dataset includes *upper-clothes* and *pants* in another dataset, and the *upper-clothes* label can be composed of more fine-grained categories (*e.g.*, *coat*, *T-shirt* and *sweater*) in the third dataset, as shown in Fig. 1. We make efforts to explore various graph transfer dependencies between different label sets, including feature-level similarity, handcraft relationship, and learnable weight matrix. Moreover, considering that the complex relationships between different semantic labels are arduous to capture from limited training data, we employ semantic similarity that is encapsulated with linguistic knowledge from word embedding [34] to preserve the semantic consistency in a scene. We encode these different types of relationships into the network to enhance the graph transfer capability.

Let $G_s = (V_s, E_s)$ denotes a source graph and $G_t = (V_t, E_t)$ denotes a target graph, where G_s and G_t may have different structures and characteristics. We can represent a graph as a matrix $Z \in \mathbb{R}^{N \times D}$, where $N = |V|$ and D is the dimension of each vertex $v \in V$. The graph transformer can be formulated as:

$$Z_t = Z_t + \sigma(A_{tr} Z_s W_{tr}), \tag{3}$$

where $A_{tr} \in \mathbb{R}^{N_t \times N_s}$ is a transfer matrix for mapping the graph representation from Z_s to Z_t. $W_{tr} \in \mathbb{R}^{D_s \times D_t}$ is a trainable weight matrix. We seek to find a better graph transfer dependency $A_{tr} = a_{i,j,\ i=[1,N_t],\ j=[1,N_s]}$, where $a_{i,j}$ means the transfer weight from the j^{th} semantic node of source graph to the i^{th} semantic node of target graph. We consider and compare four schemes for the transfer matrix.

Handcraft relation. Considering the inherent correlation between two semantic parts, we first define the relation matrix as a hard weight, *i.e.*, $\{0, 1\}$. When two nodes have a subordinate relationship, the value of edge between them is 1, else is 0. For example, *hair* is a part of *head*, so the edge value between *hair* node of the target graph and the *head* node of the source graph is 1.

Learnable matrix. In this way, we randomly initialize

the transfer matrix A_{tr}, which can be learned with the whole network during training.

Feature similarity. The transfer matrix can also be dynamically established by computing the similarity between the source graph nodes and target graph nodes, which have encoded high-level semantic information. The transfer weight $a_{i,j}$ can be calculated as:

$$a_{i,j} = \frac{exp(sim(v_i^s, v_j^t))}{\sum_j exp(sim(v_i^s, v_j^t))}, \tag{4}$$

where $sim(x, y)$ is the cosine similarity between x and y. v_i^s is the features of the i^{th} target node, and v_j^t is the features of the j^{th} source node.

Semantic similarity. Besides the visual information, we further explore the linguistic knowledge to construct the transfer matrix. We use the word2vec model [34] to map the semantic word of labels to a word embedding vector. Then we compute the similarity between the nodes of the source graph V_s and the nodes of the target graph V_t, which can be formulated as:

$$a_{i,j} = \frac{exp(s_{ij})}{\sum_j exp(s_{ij})}, \tag{5}$$

where s_{ij} means the cosine similarity between the word embedding vectors of i^{th} target node and j^{th} source node.

With the well-defined transfer matrix, the target graph features and source graph knowledge can be combined and propagated again by graph reasoning, the same as the Eq. 3. Furthermore, the direction of the transfer is flexible, that is, two graphs can be jointly transferred from each other. Accordingly, the hierarchical information of different label sets can be associated and propagated via the cooperation of Intra-Graph Reasoning and Inter-Graph Transfer, which enables the whole network to generate more discriminative features to perform fine-grained pixel-wise classification.

3.3. Universal Human Parsing

As shown in Fig. 2, apart from improving the performance of one model by utilizing the information transferred from other graphs, our Graphonomy can also be naturally used to train a universal human parsing task for combining diverse parsing datasets. As different datasets have large label discrepancy, previous parsing works must tune highly-specific models for each dataset or perform multi-task learning with several independent branches where each of them handles one level of the tasks. By contrast, with the proposed Intra-Graph Reasoning and Inter-Graph Transfer, our Graphonomy is able to alleviate the label discrepancy issues and stabilize the parameter optimization during joint training in an end-to-end way.

Another merit of our Graphonomy is the ability to extend the model capacity in an online way. Benefiting from the usage of graph transfer learning and joint training strategy, we can dynamically add and prune semantic labels for different purposes (*e.g.*, adding more dataset) while keeping the network structure and previously learned parameters.

4. Experiments

In this section, we first introduce implementation details and related datasets. Then, we report quantitative comparisons with several state-of-the-art methods. Furthermore, we conduct ablation studies to validate the effectiveness of each main component of our Graphonomy and present some qualitative results for the perceptual comparison.

4.1. Experimental Settings

Implementation Details We use the basic structure and network settings provided by DeepLab v3+ [3]. Following [3], we employ the Xception [7] pre-trained on COCO [31] as our network backbone and $output\ stride = 16$. The number of nodes in the graph is set according to the number of categories of the datasets, *i.e.*, $N = 7$ for Pascal-Person-Part dataset, $N = 18$ for ATR dataset, $N = 20$ for CIHP dataset. The feature dimension D of each semantic node is 128. The Intra-Graph Reasoning module has three graph convolution layers with ReLU activate function. For Inter-Graph Transfer, we use the pre-trained model on source dataset and randomly initialize the weight of the target graph. Then we perform end-to-end joint training for the whole network on the target dataset.

During training, the 512x512 inputs are randomly resized between 0.5 and 2, cropped and flipped from the images. The initial learning rate is 0.007. Following [3], we employ a "ploy" learning rate policy. We adopt SGD optimizer with $momentum = 0.9$ and weight decay of $5e-4$. To stabilize the predictions, we perform inference by averaging results of left-right flipped images and multi-scale inputs with the scale from 0.50 to 1.75 in increments of 0.25.

Our method is implemented by extending the Pytorch framework [33] and we reproduce DeepLab v3+ [3] following all the settings in its paper. All networks are trained on four TITAN XP GPUs. Due to the GPU memory limitation, the batch size is set to be 12. For each dataset, we train all models at the same settings for 100 epochs for the good convergence. To stabilize the inference, the resolution of every input is consistent with the original image. The code and models are available at `https://github.com/Gaoyiminggithub/Graphonomy`.

Dataset and Evaluation Metric We evaluate the performance of our Graphonomy on three human parsing datasets with different label definition and annotations, including PASCAL-Person-Part dataset [6], ATR dataset [28], and Crowd Instance-Level Human Parsing (CIHP) dataset [13]. The part labels among them are hierarchically correlated and the label granularity is from coarse to fine. Referring

Method	Mean IoU(%)
LIP [14]	59.36
Structure-evolving LSTM [24]	63.57
DeepLab v2 [2]	64.94
Li *et al*. [22]	66.3
Fang *et al*. [12]	67.60
PGN [13]	68.4
RefineNet [30]	68.6
Bilinski *et al*. [1]	68.6
DeepLab v3+ [3]	67.84
Multi-task Learning	68.13
Graphonomy (CIHP)	**71.14**
Graphonomy (Universal Human Parsing)	69.12

Table 1. Comparison of human parsing performance with several state-of-the-art methods on PASCAL-Person-Part dataset [6].

Method	Overall accuracy (%)	F-1 score (%)
LG-LSTM [27]	97.66	86.94
Graph LSTM [26]	98.14	89.75
Structure-evolving LSTM [24]	98.30	90.85
DeepLab v3+ [3]	97.30	84.50
Multi-task Learning	97.40	90.16
Graphonomy (PASCAL)	**98.32**	**90.89**
Graphonomy (Universal Human Parsing)	97.69	90.16

Table 2. Human parsing results on ATR dataset [28].

Method	Mean accuracy(%)	Mean IoU(%)
PGN [13]	64.22	55.80
DeepLab v3+ [3]	65.06	57.13
Multi-task Learning	65.27	57.35
Graphonomy (PASCAL)	**66.65**	**58.58**
Graphonomy (Universal Human Parsing)	65.73	57.78

Table 3. Performance comparison with state-of-the-art methods on CIHP dataset [13].

to their dataset papers, we use the evaluation metrics including accuracy, the standard intersection over union (IoU) criterion, and average F-1 score.

4.2. Comparison with state-of-the-arts

PASCAL-Person-Part dataset [6] is a set of additional annotations for PASCAL-VOC-2010 [11]. It goes beyond the original PASCAL object detection task by providing pixel-wise labels for six human body parts, *i.e*., *head*, *torso*, *upper-arms*, *lower-arms*, *upper-legs*, *lower-legs*. There are 3,535 annotated images in the dataset, which is split into separate training set containing 1,717 images and test set containing 1,818 images.

We report the human parsing results compared with the state-of-the-art methods in Table 1. "Graphonomy (CIHP)" is the method that transfers the semantic graph constructed on the CIHP dataset to enhance the graph representation on the PASCAL-Person-Part dataset. Some previous methods achieve high performance with over 68% Mean IoU, thanks to the wiper or deeper architecture [1, 30], and multi-task learning [13]. Although our basic network (DeepLab v3+ [3]) is not the best, the performance is improved by our graph transfer leaning, which explicitly incorporates human knowledge and label taxonomy into intermediate graph representation, then propagates and updates the global information driven by hierarchical label correlation.

ATR dataset [28] aims to predict every pixel with 18 labels: *face*, *sunglass*, *hat*, *scarf*, *hair*, *upper-clothes*, *left-arm*, *right-arm*, *belt*, *pants*, *left-leg*, *right-leg*, *skirt*, *left-shoe*, *right-shoe*, *bag* and *dress*. Totally, 17,700 images are included in the dataset, with 16,000 for training, 1,000 for testing and 700 for validation.

We report the human parsing results on ATR dataset compared with the state-of-the-art methods in Table 2. "Graphonomy (PASCAL)" denotes the method that transfer the high-level graph representation on PASCAL-Person-Part dataset to enrich the semantic information. Some previous works [24, 26, 27] use the LSTM architecture to improve the performance. Instead, we use the graph structure to propagate and update the high-level information. The advanced results demonstrate that our Graphonomy has stronger capability to learn and enhance the feature representations.

CIHP dataset [13] is a new large-scale benchmark for human parsing task, including 38,280 images with pixel-wise annotations on 19 semantic part labels. The images are collected from the real-world scenarios, containing persons appearing with challenging poses and viewpoints, heavy occlusions, and in a wide range of resolutions. Following the benchmark, we use 28,280 images for training, 5,000 images for validation and 5,000 images for testing.

The human parsing results evaluated on CIHP dataset is reported in Table 3. The previous work [13] achieve high performance with 55% Mean IoU in this challenging dataset by using multi-task learning. Our Graphonomy (PASCAL) improves the results up to 58.58%, which demonstrates its superiority and capability to takes full advantages of semantic information to boost the human parsing performance.

4.3. Universal Human Parsing

To sufficiently utilize all human parsing resources and unify label annotations from different domains or at various levels of granularity, we train a universal human parsing model to unify all kinds of label annotations from different resources and tackle different levels of human parsing, which is denoted as "Graphonomy (Universal Human Parsing)". We combine all training samples from three datasets and select images from the same dataset to construct one batch at each step. As reported in Table 1, 2, 3, our method achieves favorable performance on all datasets. We also compare our Graphonomy with multi-task learning method by appending three parallel branches upon the backbone with each branch predicting the labels of one dataset respectively. Superior to multi-task learning, our Graphonomy is able to distill universal semantic graph representation and enhance individualized representation for each label graph.

#	Basic network [3]	Adjacency matrix A^e	Intra-Graph Reasoning	Pre-trained on CIHP	Inter-Graph Transfer				Mean IoU(%)
					Handcraft relation	Learnable matrix	Feature similarity	Semantic similarity	
1	✓	-	-	-	-	-	-	-	67.84
2	✓	-	✓	-	-	-	-	-	67.89
3	✓	✓	✓	-	-	-	-	-	68.34
4	✓	-	-	✓	-	-	-	-	70.33
5	✓	✓	✓	✓	-	-	-	-	70.47
6	✓	✓	✓	✓	✓	-	-	-	70.22
7	✓	✓	✓	✓	-	✓	-	-	70.94
8	✓	✓	✓	✓	-	-	✓	-	71.05
9	✓	✓	✓	✓	-	-	-	✓	70.95
10	✓	✓	✓	✓	-	-	✓	✓	**71.14**
11	✓	✓	✓	✓	-	✓	✓	✓	70.87
12	✓	✓	✓	✓	✓	✓	✓	✓	70.69

Table 4. Ablation experiments on on PASCAL-Person-Part dataset [6].

training data	Fine-tune	Graphonomy
50%	68.45	70.03
80%	70.02	70.26
100%	70.33	**71.14**

Table 5. Evaluation results of our Graphonomy when training on different number of data on PASCAL-Person-Part dataset [6], in terms of Mean IoU(%).

Figure 4. Examples of different levels of human parsing results generated by our universal human parsing agent, Graphonomy.

We also present the qualitative universal human parsing results in Fig. 4. Our Graphonomy is able to generate precise and fine-grained results for different levels of human parsing tasks by distilling universal semantic graph representation to each specific task, which further verifies the rationality of our Graphonomy based on the assumption that incorporating hierarchical graph transfer learning upon the deep convolutional networks can capture the critical information across the datasets to achieve good capability in universal human parsing.

4.4. Ablation Studies

We further discuss and validate the effectiveness of the main components of our Graphonomy on PASCAL-Person-Part dataset [6].

Intra-Graph Reasoning. As reported in Table 4, by encoding human body structure information to enhance the semantic graph representation and propagation, our Intra-Graph Reasoning acquires 0.50% improvements compared with the basic network (#1 vs #3). To validate the significance of adjacency matrix A^e, which is defined according to the connectivity between human body parts and enables the semantic messages propagation, we compare our methods with and without A^e (#2 vs #3). The comparison result shows that the human prior knowledge makes a larger contribution than the extra network parameters brought by the graph convolutions.

Inter-Graph Transfer. To utilize the annotated data from other datasets, previous human parsing methods must be pre-trained on the other dataset and fine-tuned on the evaluation dataset, as the #4 result in Table 4. Our Graphonomy provides a Inter-Graph Transfer module for better cross-domain information sharing. We further compare the results of difference graph transfer dependencies introduced in Section 3.2, to find out the best transfer matrix to enhance graph representations. Interestingly, it is observed that transferring according to handcraft relation (#6) diminishes the performance and the feature similarity (#8) is the most powerful dependency. It is reasonable that the label discrepancy of multiple levels of human parsing tasks cannot be solved by simply defining the relation manually and the hierarchical relationship encoded by the feature similarity and semantic similarity is more reliable for information transferring. Moreover, we compare the results of different combinations of the transfer methods, which bring in a little more improvement. In our Graphonomy, we combine feature similarity and semantic similarity for the Inter-Graph Transfer, as more combinations cannot contribute to more improvements.

Different number of traning data. Exploiting the intrinsic relations of semantic labels and incorporating hierarchical graph transfer learning upon the conventional human parsing network, our Graphonomy not noly tackle multiple levels of human praing tasks, but also alleviate the need of heavy annotated traning data to achieve the desired performance. We conduct extensive experiments on transferring

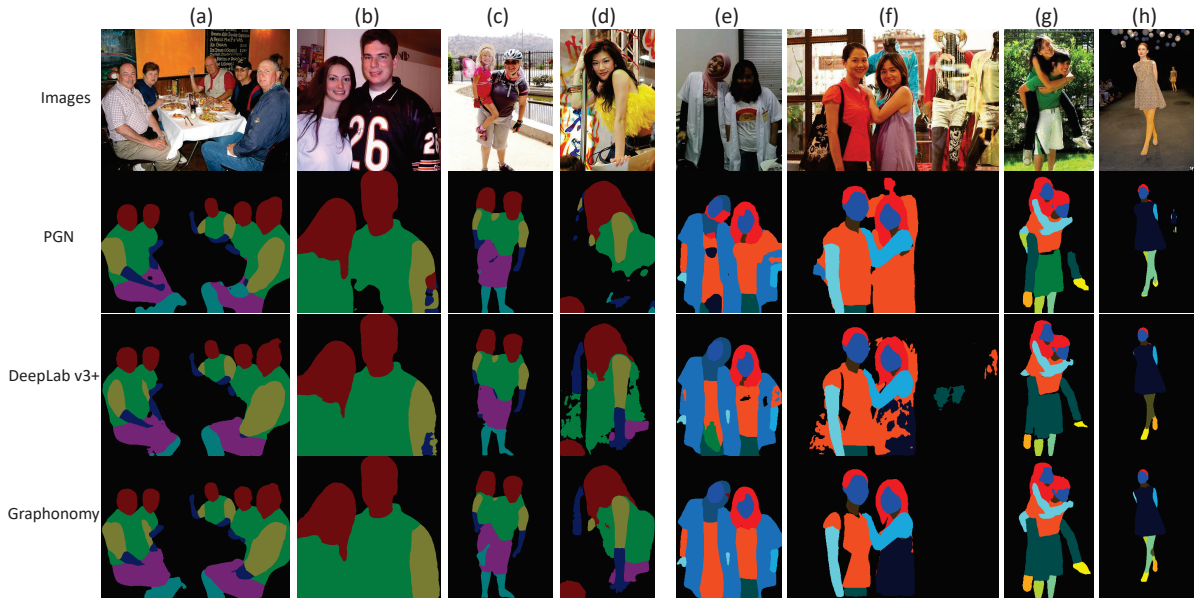

Figure 5. Visualized comparison of human parsing results on PASCAL-Person-Part dataset [6] (Left) and CIHP dataset [13] (Right).

the model pre-trained on CIHP dataset to PASCAL-Person-Part dataset. We use different annotated data in training set by random sampling for training and evaluate the models on the whole test set. As summarized in Table 5, simply fine-tuning the pre-trained model without our proposed Inter-Graph Transfer obtains 70.33% mean IoU with all training data. However, our complete Graphonomy architecture uses only 50% of the training data and achieves comparable performance. With 100% training data, our approach can even outperforms the fine-tuning baseline for 0.81% in average IoU. This superior performance confirms the effectiveness of our Graphonomy that seamlessly bridges all semantic labels from different datasets and attains the best utilization of data annotations.

4.5. Qualitative Results

The qualitative results on the PASCAL-Person-Part dataset [6] and the CIHP dataset [13] are visualized in Fig. 5. As can be observed, our approach outputs more semantically meaningful and precise predictions than other two methods despite the existence of large appearance and position variations. Taking (b) and (e) for example, when parsing the clothes, other methods are suffered from strange fashion style and the big logo on the clothes, which leads to incorrect predictions for some small regions. However, thanks to the effective semantic information propagation by graph reasoning and transferring, our Graphonomy successfully segments out the large clothes regions. More superiorly, with the help of the compact high-level graph representation integrated from different sources, our method generates more robust results and gets rid of the disturbance from the occlusion and background, like (c) and (d). Besides, we also present some failure cases (g) and (h), and find that the overlapped parts and the very small persons cannot be predicted precisely, which indicates more knowledge is desired to be incorporated into our graph structure to tackle the challenging cases.

5. Conclusion

In this work, we move forward to resolve all levels of human parsing tasks using a universal model to alleviate the label discrepancy and utilize the data annotation. We proposed a new universal human parsing agent, named as Graphonomy, that incorporates hierarchical graph transfer learning upon the conventional parsing network to predict all labels in one system without piling up the complexity. The solid and consistent human parsing improvements of our Graphonomy on all datasets demonstrates the superiority of our proposed method. The advantageous universal human parsing performance further confirms that our Graphonomy is strong enough to unify all kinds of label annotations from different resources and tackle different levels of human parsing needs. In future, we plan to generalize Graphonomy to more general semantic segmentation tasks and investigate how to embed more complex semantic relationships naturally into the network design.

6. Acknowledgements

This work was supported by the Sun Yat-sen University Start-up Foundation Under Grant No. 76160-18841201, in part by the National Key Research and Development Program of China under Grant No. 2018YFC0830103, in part by National High Level Talents Special Support Plan (Ten Thousand Talents Program), and in part by National Natural Science Foundation of China (NSFC) under Grant No. 61622214, and 61836012.

References

[1] Piotr Bilinski and Victor Prisacariu. Dense decoder shortcut connections for single-pass semantic segmentation. In *CVPR*, June 2018.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018.

[4] Xinlei Chen and Abhinav Gupta. Spatial memory for context reasoning in object detection. In *ICCV*, Oct 2017.

[5] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *CVPR*, June 2018.

[6] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, et al. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014.

[7] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, July 2017.

[8] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016.

[9] Jian Dong, Qiang Chen, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Towards unified human parsing and pose estimation. In *CVPR*, 2014.

[10] Nikita Dvornik, Konstantin Shmelkov, Julien Mairal, and Cordelia Schmid. Blitznet: A real-time deep network for scene understanding. In *ICCV*, Oct 2017.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[12] Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. In *CVPR*, June 2018.

[13] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *ECCV*, September 2018.

[14] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[17] Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. Lsda: Large scale detection through adaptation. In *NIPS*, 2014.

[18] Ronghang Hu, Piotr Dollr, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *CVPR*, June 2018.

[19] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[20] Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Multi-label zero-shot learning with structured knowledge graphs. In *CVPR*, June 2018.

[21] Jianan Li, Yunchao Wei, Xiaodan Liang, Jian Dong, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Attentive contexts for object detection. *IEEE Transactions on Multimedia*, 19(5):944–954, 2017.

[22] Qizhu Li, Anurag Arnab, and Philip HS Torr. Holistic, instance-level human parsing. *arXiv preprint arXiv:1709.03612*, 2017.

[23] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *TAPAMI*, 2018.

[24] Xiaodan Liang, Liang Lin, Xiaohui Shen, Jiashi Feng, Shuicheng Yan, and Eric P. Xing. Interpretable structure-evolving lstm. In *CVPR*, 2017.

[25] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. Deep human parsing with active template regression. *TPAMI*, 2015.

[26] Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with graph lstm. In *ECCV*, 2016.

[27] Xiaodan Liang, Xiaohui Shen, Donglai Xiang, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with local-global long short-term memory. In *CVPR*, 2016.

[28] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, 2015.

[29] Xiaodan Liang, Hongfei Zhou, and Eric Xing. Dynamic-structured semantic propagation network. In *CVPR*, June 2018.

[30] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017.

[31] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[32] Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[34] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[35] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, June 2018.

[36] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L. Yuille. Joint multi-person pose estimation and semantic part segmentation. In *CVPR*, July 2017.

[37] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. *arXiv preprint arXiv:1807.10221*, 2018.

[38] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[39] Jian Zhao, Jianshu Li, Yu Cheng, Li Zhou, Terence Sim, Shuicheng Yan, and Jiashi Feng. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. *arXiv preprint arXiv:1804.03287*, 2018.